Assignment VI : Dimensionality Reduction

April 4, 2025

ID5002W: Industrial AI Laboratory

Total Marks: 80

Instructions

1. Assignment shall be submitted before the due date. Late submissions will not be entertained. If you cannot submit the assignment due to some reasons, please contact Dr. Tirthankar by email.
2. All the assignments must be the student's own work. The students are encouraged to discuss or consult friends or classmates. However, they have to submit their own work. Any malpractice will be reported to the authorities and action will be taken as per the IIT Madras rules.
3. If you find the solution in the book or article or on the website, please indicate the reference in the solution.
4. You are expected to submit jupyter/ipython notebook as answer.
5. A short report in pdf format along with the code should be submitted containing results and analysis as asked in the assignment.
6. Please note:
   - Code should execute without any error.
   - The code should be clean with readable comments.
   - The output of the code should be in a relevant format so that it can be understood by an evaluator.
7. Grading Policy:
   - 50% code correctness.
   - 10% code readability and comments.
   - 40% on report and analysis.

Problem 1

You have been provided with a single cell RNA Sequencing dataset with expression values for different genes in different cells. You can read the data using the `read_parquet` function from pandas library. There are two category columns named `BA` and `SubGroup`.

(a) Filter the low variance genes. You can remove 10% of genes with lowest expression value. [**Marks: 10**]

(b) Apply PCA on the dataset with appropriate preprocessing. How much variance is captured by the top 100 components? [**Marks: 10**]

(c) Plot the first two components coloring based on both BA and SubGroup categories. [**Marks: 10**]

(d) Apply NMF on the dataset and extract 5 components from the dataset. Plot each of the components against one another. Identify the top 5 contributing features to each of the components. [**Marks: 20**]

(e) Apply tSNE on the data and plot the results coloring it based on `BA`. Repeat the exercise for three times with identical parameters and report if there are any changes in plot. [**Marks: 15**]

(f) Apply UMAP on the data and plot the results, coloring it based on `SubGroup`. Repeat the exercise for `n_neighbours` parameters: 8, 16 and 32. [**Marks: 15**]